# PATENT ABSTRACTS OF JAPAN

(11) Publication number : 03-194653
(43) Date of publication of application : 26.08.1991

(51) Int.Cl. G06F 15/40

(21) Application number : 01-332591
(22) Date of filing : 25.12.1989

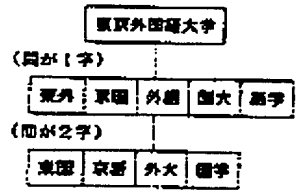
(71) Applicant : TOKAI TV HOSO KK
(72) Inventor : KASUGAI MIKIZO

## (54) METHOD FOR RETRIEVING ABBREVIATED WORD IN INFORMATION RETRIEVAL SYSTEM

(57) Abstract:

PURPOSE: To prevent a trouble in registration work by generating and registering a character index as to respective characters included in data and a phrase index as to a direct phrase consisting of two continuous characters and an indirect phrase consisting of two characters, which is combined by abbreviating one or two sandwiched words and which are combined.

CONSTITUTION: Five kinds of combinations of characters obtained by abbreviating one sandwiched character, namely, 'TO GAI/KYO KOKU/GAI GO/KOKU DAI/GO GAKU', four kinds of combinations obtained by abbreviating two sandwiched characters, namely, 'TO KOKU/KYO GO/GAI DAI/KOKU GAKU', and total nine kinds of indirect phrases are extracted. Then, the data number designation bit concerned of the key concerned of the indirect phrase index is turned on and registration for the indirect phrase index is executed. Consequently, 'TO GAI', 'GAI GO', and 'GAI DAI' are automatically registered in the phrase index as the KANJI (Chinese character) abbreviated words of 'Tokyo Gaikokugo Daigaku (Tokyo University of Foreigh Studies)'.



LEGAL STATUS

[Date of request for examination]

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

⑲日本国特許庁(JP)　⑪特許出願公開

⑫公開特許公報(A)　平3−194653

| ⑤1Int. Cl.[5] | 識別記号 | 庁内整理番号 | ㊸公開　平成3年(1991)8月26日 |
|---|---|---|---|
| G 06 F 15/40 | 510 M | 7218−5B | |

審査請求　未請求　請求項の数 1　(全5頁)

㊹発明の名称　情報検索システムにおける略語検索法

㉑特　願　平1−332591
㉒出　願　平1(1989)12月25日

㉒発明者　春日井　幹三　愛知県名古屋市東区東桜1丁目14番27号　東海テレビ放送株式会社内
㉑出願人　東海テレビ放送株式会社　愛知県名古屋市東区東桜1丁目14番27号
㉔代理人　弁理士　山下　穣平　外1名

## 明　細　書

### 1．発明の名称

情報検索システムにおける略語検索法

### 2．特許請求の範囲

情報検索システムにおいて、

データ登録時には、該データ中に含まれる各文字についての文字索引と、連続する2文字から成る直接連語、及び間に挟まれる1字あるいは2字の文字を省略して組み合わされる2文字から成る間接連語についての連語索引とを作成しておき、

データ検索時には、検索条件として指定された文字列中に含まれる各文字と前記直接連語を抽出し、これにより前記文字索引と連語索引とを検索することにより、前記検索条件の文字列を含むデータ及び、前記文字列を略語とする可能性のある文字列を含むデータを同時に得ることを特徴とする情報検索システムにおける略語検索法。

### 3．発明の詳細な説明

(産業上の利用分野)

本発明は日本語等による多量の文字情報から成るデータベースの中から検索条件に適合するデータを検索する情報検索システムに関し、特に漢字等の略語による検索を自動的に容易に行う方法に関する。

(従来の技術)

従来の情報検索システムでは、例えば「国際通貨基金」の同意語としての「IMF」等を登録するときには、別に同意語ファイルを設けて登録している。

また、例えば「東京外国語大学」の漢字略語としての「東外大」等も、本来の同意語と区別せずに、前述の同意語ファイルに登録するか、あるいは、漢字略語「東外大」を含むデータに、既にシソーラスに登録してある「東京外国語大学」のようなフルネームを添加している。

(発明が解決しようとする課題)

このような従来方式においては、漢字略語を含むデータの登録時に、同意語ファイルへの登録を怠れば検索されないため、多様な漢字略語を、すべて事前に登録しておかなければならない。例えば「東京外国語大学」の略語としては、「東京外語大」「東京外大」「東外大」等が用いられているが、これらのすべてを事前に同意語ファイルに登録しておくか、あるいはシソーラスに登録してある「東京外国語大学」のようなフルネームを添加しなければならず、いずれにしても登録作業に手間がかかるという問題点がある。

（課題を解決するための手段および作用）

本発明は、前述した課題を解決するための手段として、

情報検索システムにおいて、データ登録時には、該データ中に含まれる各文字についての文字索引と、連続する２文字から成る直接連語、及び間に挟まれる１字あるいは２字の文字を省略して組み合わされる２文字から成る間接連語についての連語索引とを作成しておき、データ検索時には、検索条件として指定された文字列中に含まれる各文字と前記直接連語を抽出し、これにより前記文字索引と連語索引とを検索することにより、前記検索条件の文字列を含むデータ及び、前記文字列を略語とする可能性のある文字列を含むデータを同時に得ることを特徴とする情報検索システムにおける略語検索法を提供するものである。

本発明の方式によれば、データ登録時に自動的に、データに含まれている各文字に関する文字索引と、直接連語に関する連語索引を登録するだけでなく、新たに、間接連語に関する連語索引を設けることによって、日本語に多い中間文字省略型の略語を同意語ファイルに登録する必要がなくなる。

また、新たに付加される間接連語に関する索引を用いることにより、通常の文字列の検索時にもより精度の高い検索が可能になる。

（実施例）

次に本発明の実施例について図面を参照して説明する。

なお、文字索引の作り方に関しては、

特願昭61-055683 「日本語情報検索システム」、

また連語索引の作り方に関しては、

特願平1-263067「情報検索システムにおける連語索引を用いた検索法」

を参照されたい。ただし、それらはあくまで１つの例であって、本発明の論理は索引の作り方に左右されるものではない。

また、連語索引を実事例に適用する場合には、主として索引の数を押さえんがために、さまざまの工夫がなされるのであるが、ここでは、論理を明らかにするために、連語索引としては、連続している２文字を扱う直接連語索引と、間に挟まれる１字あるいは２字を省略して組み合わされる２文字を扱う間接連語索引の２種類が存在しているとして説明する。

第１図は、例としてデータ「東京外国語大学」をとりあげて本実施例におけるデータ格納時の索引登録の様子を説明するためのものである。

いま、データ「東京外国語大学」が入力され、データ番号ｘを付与されてデータ部（不図示）に格納されたとする。

このとき、データに含まれているすべての文字すなわち「東／京／外／国／語／大／学」の７種が抽出され、文字索引の該当するキーの該当するデータ番号指定ビットがオンとされて、文字索引への登録が行われる（第１図(a)）。

なお、前述した様に、本発明においては、データ番号指定ビット等の索引の作り方に関しては重要ではないため、詳述は略す。第１図では点線で囲まれた文字が、上述の様にして索引に登録されたことを示している。

つづいて、すべての連続する文字と文字の組み合わせ、すなわち、「東京／京外／外国／国語／語大／大学」の６種の直接連語が抽出され、直接連語索引の該当するキーの該当するデータ番号指定ビットがオンとされて、直接連語索引への登録が行われる（第１図(b)）。

次に、間に挟まっている1字を省略してできる文字と文字の組み合わせ、すなわち、「東外／京国／外語／国大／語学」の5種、および、間に挟まっている2字を省略してできる組み合わせ、すなわち、「東国／京語／外大／国学」の4種、計9種の間接連語が抽出され、間接連語索引の該当するキーの該当するデータ番号指定ビットがオンとされて、間接連語索引への登録が行われる（第1図(c)）。

このように、本発明の方法によれば、「東京外国語大学」の漢字略語として「東外」・「外語」・「外大」などが連語索引に自動的に登録され、従来のように同意語ファイルにこれらの漢字略語を登録する手間がかからなくなる。

次に第2図は、例として漢字略語「東外大」をとりあげて、本実施例における検索時の動作を説明するための図である。

「東京外国語大学」は、このフルネームの他にも、「東京外国語大／東京外語大学／東京外語大／東京外大／東外大」などの略称でも呼ばれるため、多量のデータが格納されている日本語データベースの中にはこれらの表記法が混在している可能性があり、しかも、検索利用者は前もってそれを知ることができないのが通常である。

いま、これらのうちのどの表記法がしてあっても検索することを目的として、検索条件として、文字列「東外大」が入力されたとする。

文字列「東外大」が入力されると、まず、検索対象である全体集合のビット列と、「東」「外」「大」という3種の文字索引のビット列の間で論理積演算が行われ、これらの3文字を含んでいる集合が得られる。

次に、文字列「東外大」から、「東外」と「外大」という2種の直接連語が抽出される（第2図(a)）のであるが、いま、「東外／東□外／東□□外」（□は任意の漢字を示す）のいずれをも含んでいないデータの集合は、先に述べた連語索引への登録法から明らかなように、

NOT（直接連語「東外」OR　間接連語「東外」）

である。

また同様に、「外大／外□大／外□□大」のいずれをも含んでいないデータの集合は、

NOT（直接連語「外大」OR　間接連語「外大」）

であるから、先に文字索引を使って得た集合からこれらの集合を差し引くことによって得られる集合は、「東京外国語大学／東京外語大学／東京外国語大／東京外語大／東京外大／東外大」のいずれかの表記がされているデータをすべて含んでいることは明らかである。

以上の論理演算式を改めて示せば、

全体集合

AND　文字「東」

AND　文字「外」

AND　文字「大」

AND（直接連語「東外」OR　間接連語「東外」）

AND（直接連語「外大」OR　間接連語「外大」）

となる（第2図(b)）。

第3図は、「東京外国語大学」を意味する各種の表記法がなされた時に、どういう索引が作成されるかを具体的に示したものである。この表からも、文字列「東外大」による略語検索が指示された時に、どのように表記されていても、「東／外／大」という3種の文字索引と「東外／外大」という2種の連語索引が作成されること、したがって、どのように表記されていても検索されることは明らかである。

ただし、これら以外のデータがすべて排除される保証はなく、例えば、「関東地区外国人大会」という文字列を含むデータも集合に含まれている。しかし、検索利用者にとっては余計なデータは読みとばせば良く、必要なデータが漏れないことのほうが大切なことは言うまでもない。

なお、このように直接連語について索引を作成するだけでなく、間接連語についても索引を作成することは、単に略語検索のためばかりでなく、本来の文字列検索のためにも極めて有効である。

例えば、文字列「東外大」を含むデータを検索しようとするとき、その論理演算式は、

全体集合

AND　文字「東」

AND　文字「外」
AND　文字「大」
AND　直接連語「東外」
AND　直接連語「外大」
AND　間接連語「東大」

となり、最後の1行の「東大」を追加することによって、偶然「東外」と「外大」という2つの文字列を互いに無関係のものとして含んでいるデータが存在する場合に、これを排除することが可能となる。

次に、間接連語索引という場合に、なぜ、間に挟まる文字を1字あるいは2字に限定するかであるが、これは通常の日本語の文章の中で使用される漢字は、多くても3字までで有意味の単位になることがほとんどであるという性質を考慮してのことである。地名を例にとれば、都道府県名では「北海道」「神奈川」「和歌山」「鹿児島」の4つが3文字、他は2文字である。また、都市名でも3文字までがほとんどで、4文字は16市あるが、その多くは、「会津若松」「近江八幡」のように国名＋地名に分解可能な構造になっていて、4字で有意味になるのは「五所川原」と「八日市場」ぐらいである。これらの例からも、意味単位が4字以上になることは稀であると言えるので、略語検索のためには、間に挟まる文字を1字あるいは2字に限定することが妥当である。これ以上にすると、例えば「東大」から「東京工業大学」が検索されてしまうし、これ以下では「北大」から「北海道大学」が検索されなくなってしまう。

なお、以上の説明は、日本語情報検索システムにおける漢字略語による検索法として行っているが、カタカナの場合も有効であるし、中国語にはいっそう適している。

（発明の効果）

本発明によれば、従来方式におけるように、データ格納時に、データに含まれているすべての漢字略語を調べて、それを同意語ファイルに登録する手間や、あるいはシソーラスに収録してあるフルネームをキーワードとして添加するなどの処理は必要でなくなるため、処理速度の向上や省力化を行うことができる。

また、略語検索のために作成される間接連語索引は、通常の文字列検索の論理演算にも利用でき、精度を一層高めることができるという効果がある。

4．図面の簡単な説明

第1図は、本発明の一実施例として、「東京外国語大学」というデータが格納されるとき、どういうキーの索引が作成されるかを示す図。

第2図は、本発明の一実施例として、「東京外国語大学」を意味するさまざまな表記がなされているデータを検索する目的で、文字列「東外大」が入力されたときに、どういうキーの索引が参照されるか、及び、検索対象である全体集合のビット列と、検索条件の文字列から抽出されたキーの索引のビット列の間で行われる論理演算式を示す図。

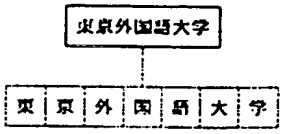
第3図は、「東京外国語大学」を意味するさまざまな表記がなされているデータについて、それぞれどういうキーの索引が作成されるか、また、「東外大」という検索条件による略語検索に際して、それらの表記法のすべてが検索されることを示す図。

代理人　弁理士　山下　穣平

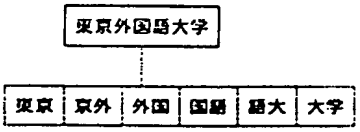
## 第 1 図

各索引への登録の実施例

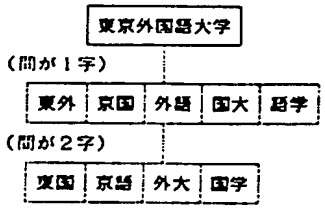
(a) (文字索引)




(b) (直接連語索引)



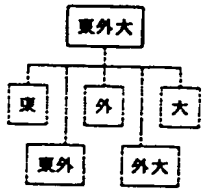

(c) (間接連語索引)



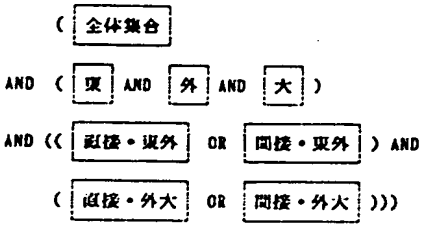

## 第 2 図

検索の実施例

(a) (検索条件)




(b)




## 第 3 図

「東京外国語大学」の
各種表記の索引キー

| 索引＼データ＼キー | 東京外国語大学 | 東京外国語大 | 東京外語大学 | 東京外語大 | 東京外大 | 東外大 | |
|---|---|---|---|---|---|---|---|
| 文字索引 東 | ○ | ○ | ○ | ○ | ○ | ○ | ← 東 |
| 文字索引 京 | ○ | ○ | ○ | ○ | ○ | ○ | |
| 文字索引 外 | ○ | ○ | ○ | ○ | ○ | ○ | ← 外 |
| 文字索引 国 | ○ | ○ | ○ | ○ | | | |
| 文字索引 語 | ○ | ○ | ○ | ○ | | | |
| 文字索引 大 | ○ | ○ | ○ | ○ | ○ | ○ | ← 大 |
| 文字索引 学 | ○ | ○ | ○ | | | | |
| 連語索引 東京 | ○ | ○ | ○ | ○ | ○ | | |
| 連語索引 京外 | ○ | ○ | ○ | ○ | ○ | | |
| 連語索引 外国 | ○ | ○ | | | | | |
| 連語索引 国語 | ○ | ○ | | | | | |
| 連語索引 語大 | ○ | ○ | ○ | ○ | | | |
| 連語索引 大学 | ○ | | ○ | | | | |
| 連語索引 東外 | △ | △ | △ | △ | △ | ○ | ← 東外 |
| 連語索引 京国 | △ | △ | | | | | |
| 連語索引 外語 | △ | △ | ○ | ○ | | | |
| 連語索引 国大 | △ | △ | | | | | |
| 連語索引 語学 | △ | | △ | | | | |
| 連語索引 東国 | △ | △ | | | | | |
| 連語索引 京語 | △ | △ | △ | △ | | | |
| 連語索引 外大 | △ | △ | △ | △ | ○ | ○ | ← 外大 |
| 連語索引 国学 | △ | | △ | △ | △ | | |
| 連語索引 東大 | | | △ | | △ | △ | |
| 連語索引 京学 | | | △ | | | | |
| 連語索引 外学 | | | △ | | | | |

○＝直接　△＝間接